Theoretical Computer Science 8 (1979) 203–225.

COMPLEXITY IN MECHANIZED HYPOTHESIS FORMATION

Pavel PUDLÁK

Mathematical Institute, Czechoslovak Academy of Sciences, Prague, Czechoslovakia

F.N. SPRINGSTEEL

University of Missouri-Columbia, Computer Science Department, Columbia, MO 65201, U.S.A.

Communicated by J. Bečvář
Received December 1976
Revised February 1978

Abstract. Practical questions of computability are studied for a special mechanized method designed to suggest scientific hypotheses on the basis of sampled data: the so-called GUHA method[1]. In simplified terms our GUHA system accepts particular data as a binary (or, binary plus "×": unknown) input matrix, which relates objects in the sample to a common set of yes-or-no properties. It seeks to output factual (non-tautologous) formal sentences, which are true in the data and so yield general hypotheses for the universe of all such objects. This paper is the first detailed analysis of the algorithmic complexity of this type of system, by considering the time (number of steps) needed to solve its basic decision problem: whether some factual sentence, of various pre-specified forms, will be so output. The resulting time bounds are functions of changeable input size and give minima for overall system complexity. In fact, when judged by the norm for efficient computability of polynomial-time, we present here both some positive and some closely related "negative" results: e.g. the distinction between *P*-time and *NP*-completeness (usually considered to be exponential time) often depends only on being given binary or ternary data, the basic question being existence of a true elementary disjunction.

Quite similar results are true for sentences with either classical or non-classical (statistically motivated) quantifiers. Moreover, some closely related two-valued problems, involving input parameters to bound desired sentence length, resisted all our efforts to place them as *P*-time *or* *NP*-complete and have an apparently intermediate complexity. At least they are concrete candidates for the (theoretical) hierarchy of *P*-reducible degrees between *P* and *NP* (assuming $P \neq NP$).

[1] The acronym "GUHA" was chosen originally to mean "General Unary Hypothesis Automaton". For the present, it should be understood merely as an artificial name for a class of methods to be described. The term "Mechanized Hypothesis Formation (MHF)" better describes the present GUHA research.

0. Introduction

Figuratively speaking, the GUHA method can be thought of as mechanically replacing that researcher's activity, who, having data about a set of objects and (some of) their properties, must decide which relations among these properties (e.g. implications) should be tested, and then which are actually true for these objects. For example, the objects are patients and their properties are diseases or facts about administration of some drugs, etc. The research worker is faced with a usually unmanageable number of possible relations to examine, even if a computer is used for the drudgery of truth validation. GUHA extends his use of the computer by automatically selecting from these relations via a thorough but heuristic generation, as well as rigorously testing them. Relevant relations are represented as formal, non-tautologous sentences, and, if discovered to be true in the data, inductively yield general hypotheses suggested for objects outside the sample. (The induction rules used for the latter are studied elsewhere; e.g. see [9].)

The clear advantage of this mechanized hypothesis formation is based on having a suitable class of formulae from mathematical logic to represent relevant relational sentences. In GUHA the classical predicate calculus is modified with respect to finite models and to generalized relations between properties, including statistical ones, to give the "observational calculi" as defined below, following Hájek and Havránek [5, 8]. More technically, properties are denoted as unary predicates and various generalized quantifiers are allowed, giving a powerful logical language of formal sentences. As input to the (general) GUHA method is a finite monadic relational structure *plus* information specifying the class of "relevant questions" (see [6, 7, 8]), the formal sentences to be investigated in the actual procedure via successive generation. The desired output is a (humanly readable) list of all "important observational statements": those relevant questions true in the model *and* logically strong (e.g., "prime" sentences, see [3, 4, 6, 7]). At least this is the final processing goal in GUHA; however, only minimal results exist about the complexity of such "full solution" methods, as in [8, Chapter 6]. Some later results of this nature will be described in [15]. Here we find the complexity of a more basic question: whether *any* sentence, of various specified forms, will be true in a given finite model. There is a fundamental connection between our results and the complexity of the full solution: clearly the latter is at least as complex as our question. Hence, the presented estimates may serve as lower complexity bounds for the full solution. The norm for "reasonable" complexity will be polynomial time.

Thus, this paper could be considered a complexity investigation of certain decision problems in finite monadic structures, independently of GUHA. However, its natural motivation lies in MHF, and we need to introduce here anyway some formal terminology from the GUHA theory. Notice that our more general structures are three-valued, with "×" added to {0, 1} to represent "unknown if true or false"; i.e., we allow for incomplete information in given models.

1. Survey of needed notions

1.1. Observational calculi

Each of these is a modification of classical predicate calculus, made not only by insisting always on finite models but by allowing a varied set of generalized quantifiers ovc predicate variables. We sketch here their definition for the single variable case; the following things specify a (monadic) *observational calculus*:

(a) Some class of finite monadic relational structures (the intended interpretations) of a certain similarity type, n. Each such structure is of the form $\langle \mathcal{M}, f_1, \ldots, f_n \rangle$ where $\mathcal{M}$ is a finite, non-void set and the f_i are functions from $\mathcal{M}$ into $\{0, 1, \times\}$, and so it can be considered to be a matrix M whose entry $M(i, j)$ is the value on the ith object of f_j. The classes which we deal with here are $\mathbf{M}_2$, the class of all $\{0, 1\}$-valued structures, and $\mathbf{M}_3$, the class of those which are $\{0, 1, \times\}$-valued.

(b) Syntactic symbols consist of:

— unary predicates $A_1, \ldots, A_n$ (over a variable, x, which can be omitted in all their occurrences).

— the classical connectives: $\&$, $\vee$, $\neg$.

— some finite set of generalized quantifiers: $Q_1, \ldots, Q_k$, of the types $s_1, \ldots, s_k$, respectively.

Formulae are defined inductively, per usual, from the atomic predicates. The induction step for quantifiers: If $\phi_1, \ldots, \phi_s$ are formulae in which x is free and Q is a quantifier of type s, then $(Qx)(\phi_1, \ldots, \phi_s)$ is a formula; Q binds x in this formula.

(c) Semantics: If ϕ is a formula in which x is free, then one defines the truth value $\|\phi\|_M[i]$, for each model M and each row i of M, in a natural way; if ϕ is a sentence (a closed formula), one defines $\|\phi\|_M$, the value of ϕ in M, to be 0 or 1 (or $\times$) in accordance with the following notions. With each connective one associates an appropriate truth table; think of the classical truth tables for connectives, in the two-valued case. For each quantifier Q, one must give an effective procedure which computes the value $\|(Qx)(\phi_1, \ldots, \phi_s)\|_M$ from the values of the functions $\|\phi_1\|_M, \ldots, \|\phi_s\|_M$ on the various rows of M. We can only be more specific by giving some examples of quantifiers which will make clear their particular truth-value definitions, as well as their evaluability in polynomial time.

1.2. Quantifier examples

(1) For calculi with two-valued structures: Let ϕ, ψ be open formulae and M a two-valued matrix. Let $m = |\{\text{rows of } M\}|$, and let r, a, b, c, d be the M-frequencies (counting rows) of the formulae $\phi, \phi \,\&\, \psi, \phi \,\&\, \neg\psi, \neg\phi \,\&\, \psi, \neg\phi \,\&\, \neg\psi$, respectively; e.g., $a = |\{i: \|\phi \,\&\, \psi\|_M[i] = 1\}|$, the number of rows which satisfy $\phi \,\&\, \psi$, etc. Note $r = a + b$. Let p be a rational number, $0 < p < 1$; e.g., $p = 0.9$. Let s be a

natural number, $s \leq m$; e.g., $s \geq 1/(1-p)$. Some quantifiers used in GUHA can be given truth definitions as follows.

Quantifiers	Type	Formula	True iff	Reading
$\forall$	1	$(\forall x)\phi$	$r = m$	for all
$\forall p$	1	$(\forall_p x)\phi$	$r \geq pm$	for relatively many
$\Rightarrow$	2	$\phi \Rightarrow \psi$	$b = 0$	implies
$\Rightarrow^+$	2	$\phi \Rightarrow^+ \psi$	$a > 0, d > 0, b = 0$	positively implies
$\Rightarrow_{p,s}$	2	$\phi \Rightarrow_{p,s} \psi$	$\frac{a}{a+b} \geq p, a \geq s$	significantly almost implies
$\sim^\circ$	2	$\phi \sim^\circ \psi$	$ad > bc$	simply deviated to

In this list only, $\Rightarrow$, and $\Rightarrow^+$ are classically definable; i.e., there is no formula in classical predicate calculus equivalent to (e.g.) $\phi \sim^\circ \psi$ in all observational structures. Furthermore, these non-classical quantifiers are meaningless for infinite structures, but they are very useful in a statistically-oriented observational calculus, forming a major part of the GUHA-MHF advantage. This was shown by Havránek in [10).

(2) For calculi with three-valued structures: First we extend the truth-tables of the connectives in such a way that 1 means "known to be true", 0 means "known to be false", and × means "no information". Thus, using the Kleene–Körner three-valued logic, the truth table of $\vee$, for example, will be:

$\vee$	0	×	1
0	0	×	1
×	×	×	1
1	1	1	1

Each quantifier defined for 2-valued structures extends uniquely to 3-valued models in the "most conservative" way. (This way has statistical motivations, and has clear practical advantages, as well.) For example, let $\sim$ represent *any* type 2 quantifier defined above for two-valued models, let F and G be two predicates, and let $M \in \mathbf{M}_3$; imagine M as a matrix with two columns. A *completion* of M is any 2-valued matrix formed from M by changing all ×'s into 0's or 1's. Then the value $\|F \sim G\|_M$ is defined to be:

1 iff $F \sim G$ is true in all completions of M,
0 iff $F \sim G$ is false in all completions of M, and
× otherwise.

We have identified only a few non-classical quantifiers, viz. those to be considered here, but there are others important in the GUHA work, towards the MHG goal. Eg., $\phi \sim_\alpha \psi$ ("is significantly deviated to") is defined by frequency conditions

$ad > bc$ *and* $\Delta(a, b, c, d) \leq \alpha$, where α is taken to be a small positive rational and $\Delta(\ )$ gives Fischer's statistic, as elaborated by Hájek and Havránek [8]. This defines $\sim_\alpha$ for 2-valued models, of course, its extension to 3-valued being as stated.

This paper presumes some minimal familiarity with and, at least, interest in the notions of algorithmic models of computations (for example, Turning machines) and of their computational complexity.

1.3. Complexity notions

There is a lot of literature about complexity of computations. We shall use notions of *P*, *NP*, polynomial reduction, *NP*-complete and an encoding of combinational problems in a way of Karp [11]. Let us only recall that here the distinction between "tractable" and "non-tractable" problem is represented by the distinction between problems in *P* and *NP*-complete problems. While a set belongs to *P* iff there is a decision procedure for it, which works in time bounded by some polynomial, the best known estimate for *NP*-complete problems is $2^{p(n)}$, for some polynomial $p(n)$. Also we shall not describe algorithms in all details in order to shorten the proofs. The reader interested in some miscellaneous problems is referred to [1, 12, 13].

We shall make use of *NP*-completeness of the following problems (for the proofs see [11]).

SATIS: Given any Boolean conjunctive normal formula μ, over some finite set of atoms and their negations, can μ be satisfied?

NODE COVER: Given any (undirected) graph $G = (X, R)$ and an integer k, is there a subset Y of X with at most k elements such that every edge in R is incident with some node in Y?

CUT: Given any (undirected) graph $G = (X, R)$, integer k, and weighting function $h : R \to N$, is there a subset Y of X such that $\sum_{y \in Y; x \notin Y} h\{x, y\} \geq k$?

Notation. Let $\mathbf{2} = \{0, 1\}$ and $\mathbf{3} = \{0, 1, \times\}$. Then $\mathbf{2}^n$ (resp., $\mathbf{3}^n$) denotes the set of all n-tuple strings over $\{0, 1\}$ (resp., over $(0, 1, \times)$). Also, $\mathbf{2}^{m \times n}$ (resp., $\mathbf{3}^{m \times n}$) denotes the set of all $m \times n$ matrices with entries from $\mathbf{2}$ (resp., $\mathbf{3}$). As earlier, $\mathbf{M}_2$ and $\mathbf{M}_3$ denote the sets of all such finite matrices.

For $M \in \mathbf{3}^{m \times n}$, let the *length* of M be $l(M) = n$, i.e., the number of its columns. For $1 \leq j \leq n$, A_j denotes the jth predicate, whose truth in M is given by the jth column: write

$$M \vDash A_j \quad \text{iff } (\forall i \leq m) M(i, j) = 1.$$

Inductively one defines $M \vDash \phi$ for other formulae ϕ, where the universal quantifier (over the implied variable, x) is usually considered to be implicit; for example, if $d = A_{j_1} \vee \cdots \vee A_{j_k}$, we write only d instead of $(\forall x)d$. Thus

$$M \vDash d \quad \text{iff } (\forall i)\{M(i, j_1) = 1 \vee \cdots \vee M(i, j_k) = 1)\}.$$

For $\alpha \in \mathbf{3}^n$, $\alpha \neq \times^n$ denote by d_α the (elementary) disjunction

$$\bigvee_{\alpha(i)=1} A_i \vee \bigvee_{\alpha(i)=0} \neg A_i$$

the *length* of d_α is to be $l(d_\alpha) = |\{i: \alpha(i) \neq \times\}|$.

Inclusion of disjunctions is defined by $d_\alpha \subseteq d_\beta$ iff $\alpha(i) = \beta(i)$ for all i with $\alpha(i) \neq \times$. Also, d_α and d_β are *disjoint* iff, for all i, $\beta(i) = \times$ whenever $\alpha(i) \neq \times$. The set of *cards* (distinct rows) of a model M is denoted by

$$C(M) = \{\alpha \in \mathbf{3}^n: \exists i\, \forall j(\alpha(j) = M(i, j))\}.$$

Note that the valuation function $\|\phi\|_M = 1$ iff $M \models \phi$ (the latter notation is preferred herein); in particular, valuation is defined for a single card, a matrix of size $1 \times n$, in exactly the same way as $\|\phi\|[i]$. For definitions of truth values of formulas with other quantifiers, see p.206. For example, $M = \phi \sim^\circ \psi$ iff $ad > bc$, where a [resp. b, c, d] is the number of cards i such that $\|\phi\|[i] = 1$ [resp. 1, 0, 0] and $\|\psi\|[i] = 1$ [resp. 0, 1, 0].

Nota Bene. All sentences considered below are (once) quantified open formulae; disjunctions are understood to be universally quantified.

2. Complexity in the two-valued case

We indicate some results, first for this case, of the computational complexity in determining whether or not there exists at least one sentence (of certain, specified forms) true in a given model. In one way, this case may seen more tractable than the three-valued. Still, the results here have great variation in complexity while the other shows mostly *NP*-completeness. Since each matrix of size $m \times n$ can be coded by input string essentially of the length $m \cdot n$, we shall relate the complexity of computation on a given matrix to this number. Thus, the answer to "Polynomial in what?" should be clear in each case below. For our purposes, also the number $\max\{m, n\}$ would suffice just as well.

There are reasons, given by practice, for which the early GUHA algorithms used only a certain set of relatively simple sentences. It appears that the elementary disjunctions bounded by various quantifiers are very suitable. This is not true for elementary conjunctions. E.g. elementary conjunctions bounded by the universal quantifier are very quickly computable, but they give little information about the model. We shall use mainly the universal quantifier $\forall$ and the quantifier of simple deviation $\sim^\circ$ as examples of classical and non-classical quantifiers. We can obtain new results by dualization. E.g. the formula $\phi \sim^\circ \psi$ is equivalent with $\neg\phi \sim^\circ \neg\psi$, hence we can replace elementary disjunctions by elementary conjunctions, here. However, questions about the existential quantifier, which is dual to the universal

quantifier, are not interesting in connection with MHF, since existential formulae are not good patterns of hypotheses.

2.1. Positive disjunctions

As an example of one of the restrictions on relevant questions in the GUHA work, one can posit the syntactical restriction that all predicates shall occur with the same signs, say without negations. A further natural restriction, for disjunctions say, is that the length be bounded by some parameter given as an "intelligibility bound", to facilitate human readability of output. While these restrictions may seem trivial, it is found that they lead to an "intractable" complexity of our first question: the existence of a positive disjunction true in a given model, the length of which is a parameter.

Definition 2.1 We say d_α is *positive*, and write $d_\alpha \in \mathrm{Pos}$, iff $\alpha \in \{1, \times\}^n$.

Theorem 2.2 *The set of pairs (M, k), consisting of a model $M \in M_2$ and a natural number k, such that there exists a positive disjunction of length at most k true in M, is NP-complete. That is, the following set has an NP-complete encoding*:

$$\mathbf{D}^+_{\mathrm{par}} = \{(M, k) \in \mathbf{M}_2 \times \mathbf{N}: \exists d_\alpha \in \mathrm{Pos},\ l(d_\alpha) \leqslant k,\ M \models d_\alpha\}.$$

Proof. Evidently $\mathbf{D}^+_{\mathrm{par}}$ is in *NP*, using an algorithm which "guesses" a k-tuple of predicates for some positive d_α and proceeds to verify it. We reduce the *NP*-complete set C_{par}, where C_{par} is the (encoded) set of (X, R, k), where (X, R) is a finite graph and k is a positive integer, for which the NODE COVER problem has a positive solution, to $\mathbf{D}^+_{\mathrm{par}}$. Consider the mapping $(X, R, k) \to (M, k)$, where M is in $\mathbf{2}^{m \times n}$ for $m = |R|$ and $n = |X|$, and the rows of M are the edges of R written as characteristic functions. That is, suppose $X = \{1, 2, \ldots, n\}$ and R is somehow ordered, say lexicographically; then

$$M(r, i) = M(r, j) = 1 \text{ iff } (i, j) \text{ is the } r\text{th edge in } R,$$

$$M(r, s) = 0 \qquad \text{otherwise, for all } 1 \leqslant r \leqslant m,\ 1 \leqslant i, j, s \leqslant n.$$

Each subset Y of X corresponds to a positive d_α with $\alpha(i) = 1$ iff i in Y, $\alpha(i) = \times$ otherwise, and conversely. Moreover, edges of R touching i in Y give rows of M in which d_α is satisfied. Therefore, $Y \subseteq X (X = R(Y)$ and $|Y| \leqslant k)$ if and only if $d_\alpha \in \mathrm{Pos} (M = d_\alpha$ and $l(d_\alpha) \leqslant k)$. Clearly an algorithm constructing (M, k) exists, the time of which is polynomial in the length, suitably encoded, of (X, R, k).

The giving of the parameter k to bound length does not affect the (seemingly) intractable complexity in this theorem. One can see this from the following corollary, where length is bounded in an intrinsic way; the same result was given by Hudlák in [14], but only via proof sketch.

Corollary 2.3. *The set of two-valued matrices M for which there exists a true positive disjunction of length at most half the length of M is NP-complete; that is,*

$$\mathbf{D}^+_{1/2} = \{M \in \mathbf{M}_2 : \exists d_\beta \in \mathrm{Pos},\ l(d_\beta) \leqslant \tfrac{1}{2} l(M),\ M \vDash d_\beta\}$$

is an NP-complete set.

Proof. Easily one can see $\mathbf{D}^+_{1/2} \in NP$. We shall reduce the *NP*-complete set $\mathbf{D}^+_{\mathrm{par}}$ to $\mathbf{D}^+_{1/2}$. Given a pair (M, k) with $M \in \mathbf{2}^{m \times n}$ and $k \leqslant n$, we transform it to a matrix $M' \in \mathbf{2}^{(m+n-k)\times 2n}$ which has these blocks:

$M' =$

$\begin{matrix} 1 & & \\ & 1 & 0 \\ & & \ddots \\ 0 & & \\ & & 1 \end{matrix}$	$\begin{matrix} \\ \\ 0 \\ \\ \\ \end{matrix}$	$\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}$
$\begin{matrix} \\ \\ 0 \\ \\ \\ \end{matrix}$	$\begin{matrix} \\ \\ M \\ \\ \\ \end{matrix}$	$\begin{matrix} \\ \\ 0 \\ \\ \\ \end{matrix}$
$n-k$	n	k

Clearly, we have $M \vDash d_\alpha$ iff $M' = d_\beta$ where $\beta = 1^{n-k}\alpha\times^k$. Thus, if some positive $d\alpha$ is true in M, then $d_\beta = A_1 \vee \cdots \vee A_{n-k} \vee d_\alpha$ is true in M'. On the other hand, for any positive d_β, $M' \vDash d_\beta$ implies that $\beta = 1^{n-k}\alpha\gamma$, where $\alpha \in \mathbf{3}^n$ and $M \vDash d_\alpha$; the arbitrary $\gamma \in \{1, \times\}^k$ is irrelevant and may as well *be* $\times^k$. Hence there is a disjunction d_α with $l(d_\alpha) \leqslant k$ true in M if and only if there is a disjunction d_β with $l(d_\beta) \leqslant n - k + k = n = \frac{1}{2} l(M')$. Evidently the case $k > n$ can be treated as $k = n$. Furthermore, the construction of M' can be done on a (multi-tape) DTM in a time polynomial in the size of M, as can be easily seen. Therefore, $\mathbf{D}^+_{\mathrm{par}} \rho \mathbf{D}^+_{1/2}$; so the latter set is also *NP*-complete, since ρ is transitive.

2.2. Elementary disjunctions

One can also investigate the existence problem, and its complexity, for true disjunctions without the syntactic restriction of positivity, only various length restrictions. Obviously one has in mind *elementary* disjunctions: those in which no predicate occurs with both signs. (In our notation d_α is elementary by necessity.) Such non-tautologous disjunctions are basic to rather many of the GUHA procedural variants, from the early work on [cf. 7]. The results obtained here reveal

either polynomial or, at least, less than exponential time complexity. The easiest disjunction problem to decide is the existence of elementary sentences of maximal length and true in a given model.

Theorem 2.4. *For any model $M \in M_2$ it can be decided in polynomial time whether there exists an elementary disjunction of maximal length true in M. [That is, $D_{\max} = \{M \in M_2: \exists d_\alpha, l(d_\alpha) = l(M), M \models d_\alpha\} \in \underline{P}$.] Moreover, there is a polynomial time algorithm which produces such a disjunction, if one exists.*

Lemma 2.5. *For $M \in \mathbf{2}^{m \times n}$ and $\alpha \in \mathbf{2}^n$, we have $M \models d_\alpha$ iff $\alpha^- \notin C(M)$, where α^- is the binary opposite of α.*

Proof. Immediate from the respective definitions.[2]

Lemma 2.6. *For $M \in \mathbf{2}^{m \times n}$, if $C(M) \subsetneq \mathbf{2}^n$,* then $\alpha \in \mathbf{2}^n \setminus C(M)$ *can be found in polynomial time.*

Proof. Consider the easily implemented algorithm which generates one-by-one the n-tuples of $\mathbf{2}^n$, ordered as binary numbers, and examines whether they equal a member of $C(M)$. After at most $m+1$ examinations, it must find an $\alpha \in \mathbf{2}^n \setminus C(M)$. Hence this algorithm works in time polynomial in $m \cdot n$, because each examination compares an n-tuple with at most the m rows in M.

Proof of Theorem 2.4. We describe an algorithm which decides if such d_α exist. Suppose $M \in \mathbf{2}^{m \times n}$ is given.

(1) Delete any repeated rows from M in order to obtain a truth-equivalent matrix $M' \in \mathbf{2}^{m' \times n}$ where $m' = |C(M)|$, by comparing in pairs the rows of M;

(2) Compare the numbers 2^n and m', in their binary forms: if $m' < 2^n$; the answer is "yes"; otherwise (i.e., if $m' = 2^n$), the answer is "no". [This is because $\exists$ such d_α with $M' = d_\alpha$ iff $\alpha^- \notin C(M') = C(M)$ iff $m' < 2^n$, by Lemma 2.5.]

Then use Lemma 2.6 to produce such d_α in the positive case. Notice that part (1) uses on the order of m^2 comparisons. In part (2) we only have to transform m' (by counting) into binary; this can also be done in at most $m^2 \geq (m')^2$ steps. Therefore, the entire procedure is polynomial in the maximum of m, n.

Corollary 2.7. *For $M \in \mathbf{M}_2$ the finding of a true elementary disjunction (if one exists) of length at most $l(M)$ can also be done in polynomial time; $\mathbf{D}_{\leq} = \{M \in \mathbf{M}_2: \exists d_\alpha, l(d_\alpha) \leq l(M), M \models d_\alpha\}$ is in P.*

[2] Observe that $\mathbf{2}^n$ can be considered as a power of the two-element group $\{0, 1\}$. We shall denote the group operation in $\mathbf{2}^n$ by $\oplus$. For $\alpha \in \mathbf{2}^n$, its *opposite is* $\alpha^- = \alpha \oplus 1^n$, where 1^n is the string of n ones.

Proof. In fact, $\mathbf{D}_{\leqslant}$ is equal to $\mathbf{D}_{\max}$, because a true d_α with $l(d_\alpha) < l(M)$ arbitrarily extends to d_β, $l(d_\beta) = l(M)$, while the maximal length ones put M in either set.

Definition 2.3. For a function $t(n)$, a language L is said to be 2^t *computable* (in time) if there is some DTM recognizing exactly L within a time bound $h^{t(n)}$, for some $h \geqslant 2$.

Thus, in view of polynomial reducibility, by allowing constant factors in powers we say that the time class $2^{\log^2}$, for example, includes all sets computable within time $h^{\log^2 n} = 2^{c \log^2 n}$, where $c = \log h$. [All logs are base 2.] This class, intermediate in complexity between P and 2^n, is important here because of the following theorem, in contrast with the result for $\mathbf{D}^+_{\text{par}}$ in Theorem 2.2.

Theorem 2.9. *The "existence of a true elementary disjunction of parametric length" set,*

$$\mathbf{D}_{\text{par}} = \{(M, k) \in M_2 \times N : \exists d_\alpha,\ l(d_\alpha) \leqslant k,\ M \models d_\alpha\},$$

is deterministically decidable in time $2^{\log^2}$.

Proof. We describe an appropriate algorithm. Suppose that $\mathbf{M} \in \mathbf{2}^{m \times n}$ and $k \in \mathbf{N}$ are given; we assume $k \leqslant n$.

(1) Compare, as binary numbers, m and 2^k. If $m < 2^k$, the answer is "yes". [Note that (M, k) is in $\mathbf{D}_{\text{par}}$ because, for any arbitrary k-tuple of predicates, say the first k, we can find an α in $\mathbf{2}^k \backslash C(M_k)$, where M_k is M restricted to just these k columns, and then d_α is true in M, by the technics of Lemmas 2.5 and 2.6 applied to M_k.]

(2) If $m \geqslant 2^k$, generate each disjunction d_α of length k (in the binary ordering of the α), and examine its truth-value in M. The number of such disjunctions is $\binom{n}{k} \cdot 2^k \leqslant n^k \cdot 2^k = (2n)^k \leqslant (2n)^{\log m} = 2^{\log 2n \log m} \leqslant 2^{\log^2(m \cdot n)}$, for $m \geqslant 2$. The verification of each d_α is polynomial, i.e., within some time $(m \cdot n)^c = 2^{c \cdot \log mn}$. Therefore, this part of the algorithm has a total bound of

$$2^{\log^2(mn)} \cdot 2^{c \log(mn)} \leqslant 2^{c' \log^2(mn)}.$$

Notice that the first part of the algorithm produces a true d_α, $l(d_\alpha) = k$, in time polynomial in $m \cdot k \leqslant m \cdot n$.

In contrast to Corollary 2.3 and to the previous theorem, we have the following "special tractability" result.

Corollary 2.10. *The median case of* $\mathbf{D}_{\text{par}}$, i.e. $\mathbf{D}_{1/2}$, *is deterministically decidable in polynomial time.*

Proof. As in the theorem, but with $k = \frac{1}{2}n$, take care of the first case ($m < 2^k$) by

answering "Yes". If the second case occurs, then $2^n \leq m^2$. [Also storage of length $O(n)$ takes only $O(\log m)$ space.] The dominating time is $\binom{n}{k} \cdot 2^k \leq 2^n \cdot 2^n \leq m^4$, where $m = \max\{m, n\}$. Therefore, the whole algorithm can be run in polynomial time [and logarithmic space], deterministically, when k is median.

One notes that a non-trivial run time happens in the $\mathbf{D}_{1/2}$ algorithm only when $m \geq \sqrt{2^n}$, i.e., when m is "very large" compared to n. This allows the trick of bounding 4^n by m^4. On the other hand, we have separate evidence that the "logarithmic case" of $\mathbf{D}_{par}:\mathbf{D}_{log} = \{M \in \mathbf{M}_2: \exists d_\alpha, l(d_\alpha) \leq \log l(M), M = d_\alpha\}$ can actually be harder than the (expected-to-be-difficult) median case, and would be happy to supply these lengthy examples to the interested reader.

Contrary to one's expected intuition is the fact that the median case of $\mathbf{D}_{par}$, where $\binom{n}{k}$ is maximized at $\binom{n}{n/2}$ is *not* its most difficult case. This is because the "harder" case $k \leq \text{MIN}\{\log n, \log m\} \leq \log(mn)$ leaves no way to bound $(2n)^k$ by a polynomial. In fact, it appears to be impossible to mimic the device in Corollary 2.3 (for positive disjunctions) in order to reduce $\mathbf{D}_{par}$ to $\mathbf{D}_{1/2}$ here; the latter is of course defined by replacing "$l(d_\alpha) \leq k$" by "$l(d_\alpha) \leq \frac{1}{2} l(M)$". Such a reduction would probably require a new, interesting method. Further, we expect that $\mathbf{D}_{par}$, $\mathbf{D}_{log}$ are *not NP*-complete problems.

If $\mathbf{D}_{par}$ were *NP*-complete, then every *NP* set would be decidable in time $2^{\log^2}$.

Hence, $\mathbf{D}_{par}$ is likely not *NP*-complete, in contrast to $\mathbf{D}^+_{par}$, because considerable work on *NP* problems has not sufficed to show them decidable, deterministically, in less than general exponential time; consult Aho et al. [1, p. 403] On the other hand, Ladner showed in [12] that if $P \neq NP$, then there are (infinitely many) degrees of polynomial reducibility between P and NP.

2.3. Non-classical sentences

Our main interest here is in the non-classical simple deviation quantifier $\sim^\circ$, which is the simplest quantifier from the class of so-called "associational quantifiers" [8, 10]. The latter include several quantifiers which are basic to the present-day GUHA algorithms because they allow practical statistical tests (e.g., Fischer's test, Chi-square test) to be used in applications. Like these others, the sentence $d \sim^\circ d'$ is *not* classically definable because there is no classical predicate calculus sentence which is truth equivalent to it in all models, as shown by Hájek and Havránek [8, Chapter 3]. It appears very possible to prove similar results about some other statistical quantifiers.

Theorem 2.11. *For the simple deviation quantifier* $\sim^\circ$, *the "existence of a true sentence of maximal length" set of (two-valued) models can be decided in poly-*

nomial time. That is, the set

$$\mathbf{D}^{\sim^\circ}_{\max} = \{M \in M_2: \exists \text{ disjoint } d_\alpha, d_\beta (l(d_\alpha) + l(d_\beta) = l(M), M \models d_\alpha \sim^\circ d_\beta)\}$$

is in P. Moreover, there is a polynomial-time algorithm which constructs such a sentence, if one exists.

Remark 2.12. In fact, this result will also hold for the (classically definable) quantifier $\Rightarrow^+$, as will be noticed during the proof. First, some definitions and lemmas.

Definition 2.13. One obtains a metric, r, on $\mathbf{2}^n$ by defining

$$r(\alpha, \beta) = |\{i: \alpha(i) \neq \beta(i)\}|.$$

For $A, B \subseteq \mathbf{2}^n$, $r(A, B) = \min\{r(\alpha, \beta): \alpha \in A, \beta \in B\}$. An easy fact: For all α, γ such that $r(\alpha, \gamma) > 1$, there is an element β between α and γ, i.e., $r(\alpha, \beta) > 0$, $r(\beta, \gamma) > 0$, and $r(\alpha, \gamma) = r(\alpha, \beta) + r(\beta, \gamma)$.

Lemma 2.14. *For $M \in \mathcal{L}^{m \times n}$ such that $C(M) \subsetneq \mathbf{2}^n$ there exist $\alpha \in C(M)$, $\gamma \in \mathbf{2}^n \backslash C(M)$ with $r(\alpha, \gamma) = 1$.*

Proof. With $C(M) \subsetneq \mathbf{2}^n$ there must exist $\alpha \in C(M)$, $\gamma \notin C(M)$ such that $r(\alpha, \gamma) = r(C(M), \mathbf{2}^n \backslash C(M))$, since $\mathbf{2}^n$ is a finite set. If $r(\alpha, \gamma) > 1$, then we have an element β between α and γ. But this gives a contradiction, for then we cannot have either $\beta \in C(M)$ or $\beta \in \mathbf{2}^n \backslash C(M)$.

Remark 2.15. We can find such α, γ in polynomial time, since we can construct the set $\{\beta: r(C(M), \beta) = 1\}$ within that bound. (Its cardinality is surely at most $\sum_{\alpha \in C(M)} |\{\beta: r(\alpha, \beta) = 1\}| = n \cdot |C(M)| \leq n \cdot m$.)

Lemma 2.16. *Suppose $M \in \mathbf{2}^{m \times n}$ has no columns of only* 1's *or only* 0's. *If $C(M) \subsetneq \mathbf{2}^n$, $n \geq 2$, then there exists a sentence ϕ true in M of the form $\phi = d_\alpha \sim^\circ d_\beta$, where d_α, d_β are disjoint, $l(d_\alpha) = 1$, $l(d_\beta) = n - 1$.*

Proof: By Lemma 2.14, take ε and δ in $C(M)$ and $\mathbf{2}^n \backslash C(M)$, respectively, such that $r(\varepsilon, \delta) = 1$. The existence of ϕ with the desired properties is not affected when we replace M by M', where $C(M') = \{\xi \oplus \pi: \xi \in C(M)\}$ for a suitable $\pi \in \mathbf{2}^n$ and if we also permute some columns, so that we can suppose $\varepsilon = 0^n \in C(M)$ and $\delta = 10^{n-1} \notin C(M)$. Since the first column also contains a 1 somewhere, there is some $\gamma \in \mathbf{2}^{n-1}$ such that $1\gamma \in C(M)$. So, $1\gamma \neq 10^{n-1}$. Set $\alpha = 1 \times^{n-1}$, $\beta = \times 1^{n-1}$; then $\phi = d_\alpha \sim^\circ d_\beta$ has the desired form. Also ϕ is true in M since we have the following table of truth values in $\mathbf{2}^n$:

ξ	Val(d_α, ξ)	Val(d_β, ξ)	M-frequencies
0^n	0	0	$d>0$
1γ	1	1	$a>0$
10^{n-1}	1	0	$b=0$

Hence, in M even $d_\alpha \Rightarrow^+ d_\beta$ is true. Finally, such α, β can be constructed for the original matrix by applying inverse transformations; ϕ can be algorithmically produced in polynomial time.

Proof of Theorem 2.1. We shall describe an algorithm which produces such a sentence, if it exists, or answers "no", if it does not exist. Let $M \in \mathbf{2}^{m\times n}$.

Step 1. Delete, for now, any columns of only 0's or only 1's from M. Let $M' \in \mathbf{2}^{m\times n'}$ be the so reduced matrix.

Step 2. Compare the binary forms of m and $2^{n'}$.

Step 3. If $m \geqslant 2^{n'}$, generate and test all possibilities of such a sentence in M'. If there exists such a ϕ true in M', go to Step 5; if not, stop; the answer is "no".

Step 4. If $m < 2^{n'}$, then there does exist a ϕ' of the desired form true in M', by Lemma 2.16, constructable in polynomial time. Find one such and continue.

Step 5. When the ith column of M contained 0's only (1's only), adjoin A_i($\neg A_i$, resp.) to one of the disjunctions of ϕ'.

Now we must show that the algorithm does the job. For this purpose it suffices to realize that, if some such maximal-length ϕ holds in M and the ith column has 0's only (1's only), then A_i must occur positively (resp. negatively) in a disjunction of ϕ, since either opposite case would make this disjunction true in all rows of M and thereby $d=0$ for ϕ, a contradiction. On the other hand, adding such vacuous A_i (or, $\neg A_i$) members does not change the validity of ϕ'.

All the steps are, in time, polynomial in $m \cdot n$; Step 3 is polynomial because the number of all (potentially) examined ϕ' is bounded by $2^{n'} \cdot 2^{n'-1} \leqslant (2^{n'})^2 \leqslant m^2$.

Remark 2.17. For the quantifier $\Rightarrow^+$, we need only modify the above algorithm at Step 3: to search for ϕ' true in M' and of the form $d_\alpha \Rightarrow^+ d_\beta$ for disjoint $\alpha, \beta (\neq \times^{n'})$ with $l(d_\alpha)+l(d_\beta)=l(M')$. The same time bounds will hold. By a trivial modification, these results also hold when "$\leqslant$" replaces "$=$". Further, for the set $\mathbf{D}^{\sim\circ}_{\text{par}}$, defined analogously to $\mathbf{D}_{\text{par}}$, one can obtain a result parallel to Theorem 2.5; here Lemmas 2.14 and 2.16 play the role in its proof played by Lemmas 2.5 and 2.6 before. Again analogous results hold for $\Rightarrow^+$, as follows. Corollary 2.10 can also be analogized.

Corollary 2.18. *The sets* $\mathbf{D}^{\Rightarrow+}_{\max}$, $\mathbf{D}^{\Rightarrow+}_{\leqslant}$, $\mathbf{D}^{\sim\circ}_{\leqslant}$, $\mathbf{D}^{\Rightarrow+}_{1/2}$ *and* $\mathbf{D}^{\sim\circ}_{1/2}$ *are in P, while* $\mathbf{D}^{\sim\circ}_{\text{par}}$ *and* $\mathbf{D}^{\Rightarrow+}_{\text{par}}$ *are* $2^{\log^2}$ *computable.*

Proof. Details, based on the preceding remark, can safely be left to the reader.

Note on implications. Just as a particular disjunction $A_1 \vee \neg A_2 \vee d$ can be viewed as an equivalent logical implication $(\neg A_1 \& A_2) \Rightarrow d$, similarly every elementary disjunction can be seen, in several forms, as an implication $c \Rightarrow d$, where c is an elementary conjunction of predicates not in d [7]. Because of this, the various logical implication problem sets—for example

$$LI_{\max} = \{M \in M_2: \text{disjoint } c, d,\ l(c)+l(d) = l(M),\ M \models (c \Rightarrow d)\},$$

are only new notation re-interpreting earlier disjunctive sets, in this case $\mathbf{D}_{\max}$. However, due to their naturality in applications, a few such (easily defined) implication sets are included in the following proposition summarizing our two-valued results.

Proposition 2.19. *For two-valued model existence problems*

(a) *the following sets are in P*:

$$\mathbf{D}_{\max},\ (LI_{\max}),\ \mathbf{D}^{\sim^\circ}_{\max},\ \mathbf{D}^{\Rightarrow^+}_{\max},\ \mathbf{D}_{\leqslant},\ (LI_{\leqslant}),\ \mathbf{D}^{\sim^\circ}_{\leqslant},\ \mathbf{D}^{\Rightarrow^+}_{\leqslant};\ \mathbf{D}_{1/2},\ (LI_{1/2}),\ \mathbf{D}^{\sim^\circ}_{1/2},\ \mathbf{D}^{\Rightarrow^+}_{1/2};$$

(b) *the following sets are* $2^{\log^2}$ *computable*:

$$\mathbf{D}_{\text{par}},\ (LI_{\text{par}}),\ \mathbf{D}^{\sim^\circ}_{\text{par}},\ \mathbf{D}^{\Rightarrow^+}_{\text{par}},\ \mathbf{D}_{\log},\ (LI_{\log});$$

(c) *the following sets are NP-complete*:

$$\mathbf{D}^{+}_{\text{par}},\ \mathbf{D}^{+}_{1/2}.$$

Problems 2.20. It would be of interest to discover the complexity of "existence of true sentence" problems for some of the more sophisticated (and statistically more important) associational quantifiers—e.g., for the Fischer-test quantifier $\sim_\alpha$. Even for the simplest associational quantifier, $\sim^\circ$, we have not discovered the status of certain problems in the two-valued case, e.g., the following, whose three-valued version is known to be *NP*-complete; cf. Theorem 3.8.

Open Problem. The complexity of

$$\mathbf{D}^{+,\sim^\circ}_{\max} = \{M \in \mathbf{M}_2: \exists \text{ disjoint } d_\alpha, d_\beta \in \text{Pos},\ l(d_\alpha)+l(d_\beta) = l(M),\ M = d_\alpha \sim_\circ d_\beta\}.$$

3. The three-valued case

We denote the three-valued versions of the previously considered problem-sets (for example, $\mathbf{D}_{\max}$, $\mathbf{D}_{\leqslant}$, $\mathbf{D}_{\text{par}}$, $\mathbf{D}^{+}_{\text{par}}$, etc.) in the following manner: $\mathbf{D}_{\max}(\times) = \{M \in \mathbf{M}_3: \exists d_\alpha (l(d_\alpha) = l(M), M \models d_\alpha)\}$. $\mathbf{D}_{\leqslant}(\times)$, $\mathbf{D}_{\text{par}}(\times)$, etc., are defined analogously.

The added possibility of uncertainty, due to entries ×, seems to make the various decision problems as hard as possible, generally. At least this is the case for all problems considered above involving the classical quantifiers, $\forall$ (over disjunctions) and $\Rightarrow$. In some cases, the two-valued versions are already known to be *NP*-complete problems, and can, as special cases, be reduced to their three-valued counterparts. Hence, the following is immediate:

The sets $\mathbf{D}^{+}_{\mathrm{par}}(\times)$ and $\mathbf{D}^{+}_{1/2}(\times)$ are *NP*-complete.

The greatest contrast between the two- and three-valued versions is in the case of elementary disjunctions: by adding uncertainty, problems which were decidable even in polynomial time previously become *NP*-complete. E.g., the following contrasts with Theorem 2.4., whose bifurcation methods do not generalize to this case.

Theorem 3.1. *Given any three-valued model M, the problem of whether there exists an elementary disjunction of maximal length true in M is NP-complete. That is, the set* $\mathbf{D}_{\max}(\times)$ *is NP-complete.*

Proof. Clearly, by guessing a d_α, $\mathbf{D}_{\max}(\times)$ is in *NP*. We show a reduction of the (*NP*-complete) satisfiable Conjunctive Normal Formula problem, SATIS, to $\mathbf{D}_{\max}(\times)$. Suppose $\mu = \&_{i=1}^{m}(D_i)$ is any CNF, where the D_i are disjunctive clauses over the set $\{j, \neg j\colon 1 \leq j \leq n\}$ of literals: n propositional atoms and their negations. Transform μ to the matrix $M \in \mathbf{3}^{m \times n}$ defined by:

$$M(i,j) = \begin{cases} 1, & \text{if } D_i \text{ contains } j, \\ 0, & \text{if } D_i \text{ contains } \neg j, \\ \times, & \text{otherwise.} \end{cases}$$

To say μ is satisfiable means that for some assignment α of truth values $\{0, 1\}$ to the atoms, using $\alpha(\neg j) = \alpha(j) \oplus 1$ and inducing over $\vee$, each clause D_i obtains the value 1. Every $\alpha : \{1, 2, \ldots, n\} \to \{0, 1\}$ is just an element of $\mathbf{2}^n$, and the corresponding d_α is an elementary disjunction over (all of) the predicates A_j of M.

We claim that some d_α is true in M iff α satisfies μ: for $i \leq m$, $\mathrm{Val}_\alpha(D_i) = 1$ iff D_i contains some literal z_j with $\alpha(z_j) = 1$; the latter occurs in two ways:

(a) if $z_j = j$, then $\alpha(j) = 1$, d_α contains A_j and $M(i,j) = 1$;

(b) if $z_j = \neg j$, then $\alpha(j) = 0$, d_α contains $\neg A_j$ and $M(i,j) = 0$.

But these are exactly the only two ways in which d_α can be true in row i of M. That is, $\|d_\alpha\|_M[i] = 1$ iff $\mathrm{Val}_\alpha(D_i) = 1$, for all i. Thus, $\mathrm{Val}_\alpha(\mu) = 1$ iff $\|d_\alpha\|_M = 1$, i.e., iff $M \models d_\alpha$.

Notice that an i,jth entry of × is needed in M in the case of both literals j and $\neg j$ being *absent* from a given clause D_i in μ. Also, the construction of M can be done in time polynomial in (viz., the square of) the larger dimension, m or n, of the

representation of μ. Therefore, $\mathbf{D}_{max}(\times)$ is also *NP*-complete, by reduction to it of its isomorphic problem, SATIS.

Example 3.2. The Boolean formula

$$\mu = (\neg 1) \& (\neg 2) \& (3 \vee 4) \& (\neg 4)$$

has as satisfying assignment only $\alpha = \langle 0, 0, 1, 0 \rangle$; its corresponding three-valued matrix M, below, has as its only true disjunction $d_\alpha = \neg A1 \vee \neg A_2 \vee A_3 \vee \neg A_4$. However, conjoining to μ the single clause $(\neg 3)$ will render the formula unsatisfiable, and its (five-by-four) matrix will have no true disjunction. (Although it would, erroneously, if any $\times$ in the fifth row were marked as a zero!)

$M =$

A_1	A_2	A_3	A_4
0	×	×	×
×	0	×	×
×	×	1	1
×	×	0	×

Observation: The contrast of this result with the solution of the two-valued $\mathbf{D}_{max}$ problem (in *P*, by Theorem 2.4) can be appreciated via the following facts. A three-valued M has a true elementary disjunction d iff d is true in the *full* two-valued completion $\tilde{M}$ formed from the union of all possible completions of rows of M. However, the difference in $C(M)$ and $C(\tilde{M})$ may be exponential: sometimes $|C(\tilde{M})| \geq 2^{n-1} \cdot |C(M)|$, where $n = l(M) = l(\tilde{M})$. In the same way, the SATIS problem for CNF's "literally full in each clause" (each D_i contains either j or $\neg j$) is decidable in *P*, via dimension comparison. But, to expand an arbitrary CNF into this full form can, again, take exponential time.

This contrast is related to the actuality that many "incomplete information" problems are hard.

Corollary 3.3. *The sets* $\mathbf{D}_{\leq}(\times)$ *and* $\mathbf{D}_{par}(\times)$ *are both NP-complete.*

Proof. The first set is in fact equal to $\mathbf{D}_{max}(\times)$, via the same idea as in Corollary 2.7; the second can have $\mathbf{D}_{max}(\times)$ reduced to it as a special case, where $k = n$.

Corollary 3.4. *The set* $\boldsymbol{D}_{1/2}(\times)=\{M\in\mathbf{M}_3\colon \exists d_\alpha,\ l(d_\alpha)\leqslant\frac{1}{2}l(M),\ M\models d_\alpha\}$ *is NP-complete.*

Proof. Clearly $\mathbf{D}_{1/2}(\times)\in NP$. Reduce $\mathbf{D}_{\leqslant}(\times)$ to it by mapping $M\in\mathbf{2}^{m\times n}$ to $M''\in\mathbf{2}^{m\times 2n}$, where $M^\circ=[M|\times]$. Only disjunctions using (some of) the first n predicates can be true in M'', since the last n are not verifiable. Hence, some d_α with $l(d_\alpha)\leqslant l(M)=n=\frac{1}{2}l(M'')$ is true in M iff it is true in M''.

Remark 3.5. For the quantifier $\sim^\circ$ (and the related one $\Rightarrow^+$), we have only a few "positive" three-valued results. One of these involves the condition $O(C(M))=2^{n-1}$, which is considered here because matrices with $n\doteq O(m)$ are not so frequent, or interesting, in GUHA applications. (Usually one has many more objects than properties!) The special problems below are shown decidable in P; in their proofs, one algorithmic phase, validity checking, has new features and requires the concept of "critical completion". A two-valued matrix M' is a *critical completion* of $\mathbf{M}\in M_3$ for the quantifier $\sim^\circ$ if the following holds: $\|F\sim^\circ G\|_M=1$ if and only if $\|F\sim^\circ G\|_{M'}=1$, i.e., iff arbitrary sentences $F\sim^\circ G$ are true in M' just when true in M. (For other quantifiers, critical completions can be defined similarly; sometimes they are non-trivial to characterize or compute.) For $\sim^\circ$ this means that M' can be characterized as the "worst case" 2-valued matrix constructed from M by substituting 0's or 1's for ×'s, in such a way that the (row-frequency) product $b\cdot c$ is made as large as possible without increasing $a\cdot d$, minimizing $ad-bc$. Let

$$\mathbf{D}^{+,\sim^\circ}_{\max-1}(\times)=\{M\in\mathbf{M}_3\colon \text{disjoint } d_\alpha, d_\beta\in\mathrm{Pos},$$
$$l(d_\alpha)=1, l(d_\beta)=l(M)-1, M\models(d_\alpha\sim^\circ d_\beta)\}.$$

$\mathbf{D}^{+,\Rightarrow^+}_{\max-1}(\times)$ is defined similarly, using $\Rightarrow^+$ in place of $\sim^\circ$. These sets are important to applications where a single property (e.g., cancer) needs to be correlated to combinations of other properties. Define the restricted set, where c is any positive constant, $\mathbf{D}^{\sim^\circ}_{c,\max}=\{M\in\mathbf{3}^{m\times n}\colon 2^{n-1}\leqslant m^c$ & disjoint $d_\alpha, d_\beta l(d_\alpha)+l(d_\beta)=l(M), M\models d_\alpha\sim^\circ d_\beta\}$. Similarly, $\mathbf{D}^{\Rightarrow^+}_{c,\max}$.

Theorem 3.6. *The following sets are in* P:

(a) $\mathbf{D}^{+,\sim^\circ}_{\max-1}(\times)$;

(b) $\mathbf{D}^{\sim^\circ}_{c,\max}(\times)$, *where* c *is a positive fixed constant. The latter gives the decidability, for models* $M\in\mathbf{M}_3$ *with* $n=2(M)\leqslant c\cdot\log|C(M)|=c\cdot\log m$, *of the existence of a true sentence of form* $d_\alpha\sim^\circ d_\beta$ *in polynomial time.*

Proof. (a) The number of possibilities for disjoint, positive (or even negative) d_α, d_β with $l(d_\alpha)=1$, $l(d_\beta)=n-1$, is linear in n. To check each such $d_\alpha\sim^\circ d_\beta$, use the critical ("worst case" substitution) completion of M for $\sim^\circ$. Significantly, this two-valued matrix has exactly as many rows as M and is straightforward to

construct even on TM tape. Thus, the obvious algorithm runs in polynomial time.

(b) Given such an M, here it generates each possibility of non-void d_α, d_β of complementary lengths. The number of choices of a non-trivial subset of n predicates, *plus* the choices of its signs, is bounded by

$$(2^n-2)\cdot 2^n \leqslant (2^{n-1}-1)\cdot 2^{n+1} \leqslant 4m^{2c}.$$

Each generated sentence can be checked in M, as above, in polynomial time. So these sets are in P. Likewise any finite union of them, for different constants, is also in P.

Corollary 3.7. *The following sets are in P*:

(a) $\mathbf{D}^{+,\Rightarrow^+}_{\max-1}(\times)$; (b) $\mathbf{D}^{+,\Rightarrow^+}_{c,\max}(\times)$; (c) $\mathbf{D}^{+,\sim^c}_{c,\max}(\times)$.

Proof. Use the same algorithmic methods as in the theorem. Even the "worst case" completion of M can be the same, in cases (a) and (b), because it attempts to increase the M-frequency of (b) (and of (c)), while (a) and (d) cannot increase.

Lastly we give some *NP*-completeness results for the quantifier $\sim^c$ in the three-valued case, which bear evidence that incomplete information makes problems "as hard as possible", under general conditions. Even some problems which are left open in the two-valued case can be resolved here, of which the following theorem is an example.

Theorem 3.8. *The set of three-valued models M, such that there are two disjoint positive disjunctions of (combined) maximal length simply deviated in M, is NP-complete. That is, the set* $\mathbf{D}^{+,\sim^c}_{\max}(\times)$ *is NP-complete.*

Proof. For a weighted (undirected) graph (X, h), where $h: P_2(X) \to \{0\} \cup \mathbf{N}$, let $w_h = \sum_{x,y\in X} h\{x, y\}$ and denote the value of any cut $(Y, X\setminus Y)$ by $W_h(Y) = \sum_{y\in Y, x\in Y} h\{x, y\}$. $P_2(X)$ is the set of all doubletons of verticles in X. Now the *NP*-complete CUT problem set is denoted by

$$\mathrm{CUT}_{\mathrm{par}} = \{(X, h, k): (X, h) \text{ a weighted graph, } k \in N,\ Y \subseteq X (W_h(Y) \geqslant k)\}.$$

We can suppose some further restrictions on (X, h, k), which do not affect the *NP*-completeness of $\mathrm{CUT}_{\mathrm{par}}$. Namely, let $h\{x, y\}$ be even for all $x, y \in X$. So we need only consider k such that $2 < k \leqslant w_h$, k even.

We will reduce $\mathrm{CUT}_{\mathrm{par}}$ to the studied set. Suppose $X = \{1, 2, \ldots, n\}$. Then the following matrix M will correspond to (X, h, k):

$$M = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \end{bmatrix} \quad \text{where } M_i \in \mathbf{3}^{m_i \times n}; \text{ each } m_i \text{ and } M_i \text{ is defined below.}$$

M_1 consists of 0's only, M_2 of crosses only, and M_3 is a "copy" of (X, h) such that for $\{i,j\}\in P_2(X)$ there are exactly $h\{i,j\}$ rows of γ in M_3, where $\gamma\in 2^n$ is the characteristic function of $\{i,j\}$ in the edge set. Hence $m_3 = w_h$, and the choices of m_1 and m_2 will be motivated below.

There is a natural $1-1$ correspondence between cuts and the considered formulae; let a cut $(Y, X\setminus Y)$ correspond to $\alpha,\beta\in\{1,\times\}^n$, $d_\alpha \sim^\circ d_\beta$, where $\alpha(i)=1$ iff $i\in Y$, and $\beta(i)=1$ iff $i\notin Y$. Consider a critical completion of M, i.e., some completion to a binary matrix N where $ad-bc$ is minimal; a, b, c, d are the frequencies occurring in the evaluation of $d_\alpha \sim^\circ d_\beta$. The crucial point here is that the frequency a in N, as well as in M_3, is equal to $W_h(Y)$, the value of the corresponding cut. Further, $d = m_1$ and $b+c = m_2+w_h-a$. Suppose $m_2\geq w_h$ and m_2+w_h is even. Under these conditions the maximum product bc is obtained for $b=c=\frac{1}{2}(m_2+w_h-a)$. Denote by $f(m_1, m_2, x)$ the function which computes the difference $ad-bc$ from m_1, m_2 and x = the value of an arbitrary cut; that is,

$$f(m_1, m_2, x)=x\cdot m_1-\tfrac{1}{2}(m_2+w_h-x))^2.$$

Now it only remains to find suitable m_1 and m_2, with $m_2\geq w_h$ and m_2+w_h even, such that for all even numbers x and k, $0\leq x\leq w_h$ and $2<k\leq w_h$, one has

(*) $\quad f(m_1, m_2, x)>0$ iff $x\geq k$.

Put

$$m_1=\tfrac{1}{4}w_h^2\cdot(k-2).$$

$$m_2=(w_h+1)(k-2)-w_h=w_h\cdot(k-2)-w_h+k-2.$$

Then $f(m_1, m_2, k-2)=0$, and, since f in increasing in $x\in[0, w_h]$, the condition (*) is satisfied. So M has such a true formula, $d_\alpha\sim^\circ d_\beta$, if and only if $ad-bc>0$ (in some critical completion), i.e., iff condition (*) holds, and it is exactly in this case that the corresponding cut satisfies $W_h(Y)\geq k$.

Corollary 3.9. *The set* $\mathbf{D}^{\sim^\circ}_{\max}(\times)$ *is also NP-complete.*

Proof. Replace 0 by $\times$ in M_3 in the proof of the theorem. Then

0
1
1

is a completion of the so-obtained M. Hence, only positive or negative disjunctions might be simply deviated in M. On the other hand, the completion of M replacing each $\times$ (in M_2 and M_3) by 0 gives a matrix in which no negative disjunctions can be

deviated, because it has no row of only 1's, so d would be 0. We ignore the trivial case when $n \leq 2$; the rest of the proof goes through as before.

Corollary 3.10. *The sets* $\mathbf{D}_{\leq}^{+,\sim^\circ}(\times)$, $\mathbf{D}_{\leq}^{\sim^\circ}(\times)$ *are NP-complete.*

Proof. It is not hard to see that the just preceding reductions carry through also for these two sets. The reason is that, just as sentences of maximal length $l(M)$ correspond to cuts of the given graph, sentences of length less than or equal to $l(M)$ correspond to cuts on subgraphs where their value always bounds from below that of coinciding cuts for the whole graph.

Corollary 3.11. *The four sets* $\mathbf{D}_{\text{par}}^{+,\sim^\circ}(\times)$, $\mathbf{D}_{1/2}^{+,\sim^\circ}(\times)$, $\mathbf{D}_{\text{par}}^{\sim^\circ}(\times)$, $\mathbf{D}_{1/2}^{\sim^\circ}(\times)$ *are NP-complete.*

Proof. Reduce $\mathbf{D}_{\max}^{+,\sim^\circ}(\times)$, $\mathbf{D}_{\max}^{\sim^\circ}(\times)$ to $\mathbf{D}_{\text{par}}^{+,\sim^\circ}(\times)$, $\mathbf{D}_{\text{par}}^{\sim^\circ}(\times)$, respectively, via the map $M \mapsto (M, l(M))$. Also, one can reduce the problems of Corollary 3.10 to the half-length problems here by the device of Corollary 3.4.

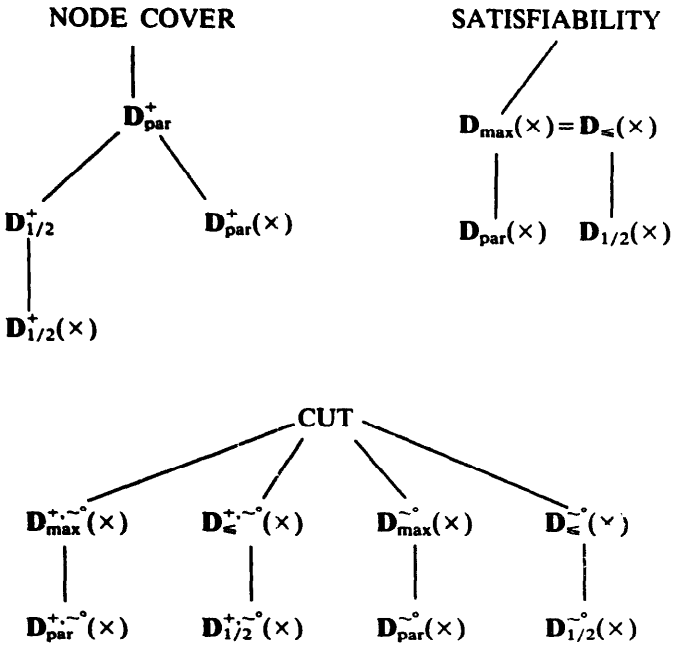
Note: The tables below summarize most of our main results, including some implicit ones. Open problems are denoted by "?", but of course all problems fall within the *NP* class. In view of the Appendix, we find no significant three-valued problems which are tractable, except ones clearly in *P*. No three-valued considered have been left open.

Table 1. Summary of results

		Logical values			
		2		3	
Restriction	Length	$\forall$	$\sim^\circ$	$\forall$	$\sim^\circ$
	parameter	$2^{\log^2}$	$2^{\log^2}$	*NP*-c.	*NP*-c.
	$\frac{1}{2}l(M)$	*P*	*P*	*NP*-c.	*NP*-c.
	$l(M)$	*P*	*P*	*NP*-c.	*NP*-c.
	$\leq l(M)$	*P*	*P*	*NP*-c.	*NP*-c.
	parameter	*NP*-c.	?	*NP*-c.	*NP*-c.
	$\frac{1}{2}l(M)$	*NP*-c.	?	*NP*-c.	*NP*-c.
positive	$l(M)$	*P*	?	*P*	*NP*-c.
	$\leq l(M)$	*P*	?	*P*	*NP*-c.

NP-c. = *NP*-complete

Table 2. Table of Reductions of *NP*-complete sets



Acknowledgement

The authors express their deep appreciation to colleague and friend Dr. Petr Hájek for introducing us to the problems here (and to each other), for reading several versions of the paper and suggesting improvements, and for general inspiration and encouragement. The essential part of this work was done while F.N. Springsteel was at the Czechoslovak Academy of Sciences, Prague (Mathematical Institute and Bio-Mathematical Center) and at the University of Kaiserslantern, F.R.G. (Fachbereich Informatik). We would also like to thank Miss Joan Cruze for her careful typing of the manuscript.

Appendix

There is an important and basic thesis in [8], which asserts that every statistical test can be considered as a quantifier in some monadic predicate calculus. We want to show that some of the presented results extend to such "statistical quantifiers".

Consider the χ^2-quantifier, $\sim^2_\alpha$, related to the χ^2-test of independence of two predicates. This quantifier is defined in terms of the frequencies a, b, c, d similarly as $\sim^\circ$:

$$\phi \sim^2_\alpha \psi \Leftrightarrow ad > bc \quad \text{and} \quad \frac{(ad-bc)^2}{rskl} m \geq \chi^2_\alpha$$

where $m=a+b+c+d$, $r=a+c$, $s=b+d$, $k=a+b$, $l=c+d$, and χ^2_α is the α-quantile of χ^2 distribution. More precisely, we define infinitely many quantifiers—one for every α. In order for the relation $\geqslant\chi^2_\alpha$ to be computable in polynomial time, we shall confine ourselves to α's such that χ^2_α is rational.

Theorem. $\mathbf{D}^Q_L(\times)$ *and* $\mathbf{D}^{+,Q}_L(\times)$ *are NP-complete for* $Q=\sim^2_\alpha$ *and* $L\in$ {par, max, $\frac{1}{2}$, $\leqslant$}.

Sketch of the proof. Suppose M is two-valued. Put

$$\mathrm{chi}(\Phi, M)=\frac{(ad-bc)^2}{rskl}m,$$

where Φ is of the form $\phi\sim^2_\alpha\psi$, and a, b, c, d are the corresponding frequencies. Then the relation between $\sim^\circ$ and $\sim^2_\alpha$ is expressed by the equivalence

$$M=\phi\sim^2_\alpha\psi\Leftrightarrow M=\phi\sim^\circ\psi \quad\text{and}\quad \mathrm{chi}(\Phi, M)\geqslant\chi^2_\alpha.$$

For a matrix M, denote by $M^{(t)}$ the following matrix

$$M^{(t)}=\begin{array}{|c|}\hline M\\ \hline \vdots\\ \hline M\\ \hline\end{array}\quad t\text{—times.}$$

Then $\mathrm{chi}(\Phi, M^{(t)})=t\cdot\mathrm{chi}(\Phi, M)$. Observe that $\mathrm{chi}(\Phi, M)>(ad-bc)^2/m^3$. For $t=m^3\chi^2_\alpha$, we have $\mathrm{chi}(\Phi, M^{(t)})\geqslant(ad-bc)^2\chi^2_\alpha$. That is $\mathrm{chi}(\Phi, M^{(t)})\geqslant\chi^2_\alpha$ whenever $ad>bc$, which means that there is no difference between $\sim^\circ$ and $\sim^2_\alpha$ in $M^{(t)}$. Hence $M\to M^{(t)}$, $t=m^3\chi^2_\alpha$ is a reduction of a problem for $\sim^\circ$ to the same problem for $\sim^2_\alpha$. Some consideration of critical completions is needed for M three-valued; we leave it to the reader. All the problems are *NP*-complete for $Q=\sim^\circ$ and are *NP* for $Q=\sim^2_\alpha$. Hence the reduction gives that they are *NP*-complete for $Q=\sim^2_\alpha$ too.

References

[1] A. Aho, J. Hopcroft and J. Ullman, *The design and Analysis of Computer Algorithms* (Addison-Wesley, Reading, MA, 1974).

[2] S.A. Cook, The complexity of theorem-proving procedures, *Proc. Third Annu. ACM Symposium on Theory of Computing* (1971) 151–158.

[3] P. Hájek, Automatic listing of important observational statements I, II, III, *Kybernetika* **9** (1973) 187–205, 251–271, *Kybernetika* **10** (1974) 95–124.

[4] P. Hájek, Some logical problems of automated research, *Proc. Symp. on Mathematical Foundations of Computer Science, High Tatras, Czechoslovakia* (1973) 85–93.

[5] P. Hájek, On logics of discovery, in *Mathematical Foundations of Computer Science 1975, Proceedings 1975*, Lecture Notes in Computer Science **32** (Springer-Verlag, Berlin, 1975) 30–45.

[6] P. Hájek, K. Bendová and Z. Renc,: The GUHA method and the three-valued logic, *Kybernetika* **7** (1971) 421–435.

[7] P. Hájek, I. Havel and M. Chytil, The GUHA method of automated hypotheses determination, *Computing* **1** (1966) 293–308.

[8] P. Hájek and T. Havránek, *Mechanized Hypothesis Formation: Mathematical Foundations for a General Theory* (Springer-Verlag, Berlin, 1978).

[9] T. Havránek, The approximation problem in computational statistics, in: *Mathematical Foundations of Computer Science 1975, Proceedings 1975*, Lecture Notes in Computer Science **32** (Springer-Verlag, Berlin, 1975) 258–265.

[10] T. Havránek, Statistical quantifiers in observational calculi: an application in GUHA methods, *Theory and Decision* **6** (1975) 213–230.

[11] R. Karp, Reducibility among combinatorial problems, in: R.E. Miller and J.W. Thatcher, Eds. *Complexity of Computer Computations* (Plenum Press, New York, 1972) 85–103.

[12] R.E. Ladner, On the structure of polynomial time reducibility, *J. Assoc. Comput. Math.* **22** (1975) 155–171.

[13] R.E. Ladner, N.A. Lynch and A.L. Selman, A comparison on polynomial time reducibilities, *Theoret. Comput. Sci.* **1** (1975) 103–123.

[14] P. Pudlák, Polynomially complete problems in the logic of automated research, in: *Mathematical Foundations of Computer Sciences 1975, Proceedings 1975*, Lecture Notes in Computer Science **32** (Springer-Verlag, Berlin, 1975) 358–361.

[15] F. Springsteel, Complexity of hypothesis-inference problems, in: P. Hájek, T. Havránek and X. Chytil, Eds., *Hypotheses, Inference and Computation* (tentative title), Theory and Decision Library (Reidel Publ., in preparation).